ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

E-MAIL SPAM DETECTION USING LAZY ASSOCIATIVE APPROACH AND NAÏVE BAYE'S FILTERING.

Asadullah Shaikh, Parkar Alifiya, Ferheen Salmani, Shiba Sayed.
Dept. of Computer Engineering, M. H. Saboo Siddik College of Engineering, Mumbai, India.

Manuscript Info

Manuscript History:

Received: 14 January 2016
Final Accepted: 26 February 2016
Published Online: March 2016

Key words:
Question Genetic Algorithm, Local Search, Heuristic Crossover, Natural Selection.

Corresponding Author

Parkar Alifiya.

Abstract

Email is one of the most modern and widely used means of communication nowadays, mainly due to its efficiency, low cost, and compatibility of diversified types of information. However, spammers are continuously crawling the Web for email addresses available at Web pages, so that more and more and more people can be reached, thus eroding away much of the attractiveness of email communication. Not only Spam frustrating for most email users, it strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. The proposed approach discovers patterns hidden in the message, and then it builds a classification model by exploring the associations among the discovered patterns.

Copy Right, IJAR, 2016,. All rights reserved.

Introduction:-

Email provides a perfect way to send millions of advertisements at no cost for sender and this is the reason mail boxes are cluttered with spam [1]. But whatever it is, spammers have to deliver their message to the user. This fact makes the message itself a weak point of the spammer. Therefore we propose a content based spam detection technique. The proposed system shall make use of naïve baye's filtering algorithm, therefore, the words in the subject will be compared with those are compared with the words stored in the database; if match is found then the mail will be discarded. Further, if a match is not found, lazy associative classification rule is applied and a classification model is developed on the basis of content of the message to find out whether the mail falls under spam or legitimate class.

Existing System:-

Lazy Associative Classification Rule:-

Classification aims to map a data instance to its appropriate class or label. In associative classification the mapping is done through an association rule with the consequent restricted to the class attribute. Eager associative classification algorithms build a single rule set during the training phase, and this rule set is used to classify all test instances. Lazy algorithms, however, do not build a rule set during the training phase; the rule set generation is delayed until a test instance is given.

❖ **Definitions:-**

[ITEMSETS]
For any set X, its size is the number of elements in X.
Let I denote the set of n natural numbers
{1, 2,. . .,n}. Each x∈ I is called an item. A non-empty subset of I is called an itemset. An itemset of size k, X={x1,x2,…xk} is called a k-itemset. We say that X contains Y if Y ⊆X.[1]**.**

❖ **Training and tests instances:-**

A data instance Ti is an itemset where I is a natural number called instance identifier or tid. A dataset D is a finite set of instances, and it is divided into two partitions,

D=Dseen ∪Dunseen, whereDseen is the set of training instances (i.e., the training data) and Dunseen is the set of test instances(i.e., the test data). If Ti is followed by the class attribute c, then Ti∈ Dseen, otherwise Ti ∈ Dunseen. The support of an itemset X is the fraction of training instances that contain X, given as,

$\sigma(X) = |\{Ti \in Dseen | X \subseteq Ti\}||Dseen|$ …(1)

The itemset X may appear too frequently in some

classes, and too rarely in others. Weighted support is the support of X normalized by the support of each class, and it is given as, by

$\gamma(X) = \sigma(X \cup c)\sigma(c)$ …(2)

❖ **Association rules:-**

An association rule is a rule with the form $X\theta, \varphi, \pi ---\rightarrow c$, (3) $X\theta, \varphi, \pi ---\rightarrow c$,

where c is the class attribute and X is an itemset (c /∈ X). The An association rule is a rule with the form Xθ,φ,π−−−→c, where c is the class attribute and X is an itemset (c /∈ X). The rule, denoted as θ, is given by,

$\sigma(X \cup c)\sigma(X)(4)$ [1]

The exclusiveness (or weighted confidence) of the rule, denoted as φ, is given by,

$$\gamma(X \cup c)\gamma(X \cup c) + \gamma(X \cup c) \ldots (5)$$

The higher the exclusion, strongly X is associated to class c. The conviction of the rule, denoted as π, is given by,

$$\sigma(X) \times \sigma(c)\sigma(X \cup c) \ldots (6)$$

and measures implication[1]. It is directional and it is maximal (π=∞) for perfect implications, and it indicates when the rule does not hold anything more than expected (π= 1). A rule

$$X\theta, \varphi, \pi ---\rightarrow c(7)$$

matches a data instance Ti if X ⊆Ti.

Bayesian Classifier:-

These types of classifier uses Bayes theorem, which states that

- probability(cj | d) = probability(d | cj) probability(cj) probability(d).
- probability(cj | d) = probability of instance d being in class cj, This is what we are trying to compute.
- probability(d | cj) = probability of generating instance d given class cj, We can imagine that being in class cj, causes you to have feature d with some probability.
- probability(cj) = probability of occurrence of class cj, This is just how frequent the class cj, is in our database.
- probability(d) = probability of instance d occurring[2]. It does not hold much importance, since it is the same for all classes.
- Assume that we have two classes c1 = male, and c2 = female.
- We have a person whose sex we do not know, say "joe" or j. Classifying drew as male or female is equivalent to asking is it more probable that d is male or female, i.e which is greater probability(male| joe) or probability(female| joe).
- probability(male| joe) = probability(joe | male) probability(male)/probability(male).

Proposed system:-

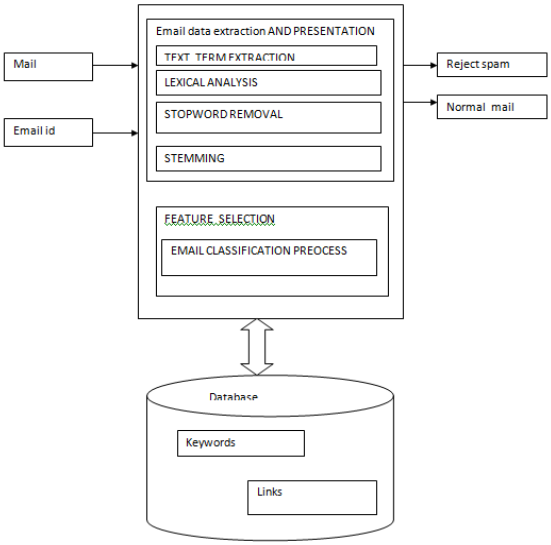

Fig: Classifier using proposed approach.

In the proposed system we make use of features of both the algorithms inorder to cause a better impact on classification process. There are two general approaches to mail filtering: knowledge engineering (KE) and machine learning (ML). In the former case, a set of rules is created according to which messages are categorized as spam or legitimate mail. [4]

Naïve Bayes is a generative classification method that is based on Bayes theorem. It calculates the prior probabilities of each class and probabilities of each attribute in each class. It assumes that the probabilities of each attribute are independent of each other. At the time of classification it uses the prior probabilities of each class and the probabilities of the observed attributes. The class with highest probability is assigned to the instance being classified.

The proposed system introduces a novel content-based spam detection approach, which first uncovers patterns hidden in both spam and legitimate messages, and then it associates the discovered patterns with the corresponding class .Not only it uses words and links for classification of message, but it uses the combination of them. [7] This approach is efficient in terms of computational complexity, being able to classify more than one hundred messages per second.

Association rules generation process is as given below, Given a set of messages D, and thresholds σmin and θmin, the task of generating association rules is to find all strong rules in D. This task can be divided in two steps:

- Enumerate frequent patterns: The set of all patterns X for which $\boldsymbol{\sigma(X) \geq \sigma min}$ ... (7) is generated.
- Generate strong rules: The set of all rules X → c for which X is frequent and $\boldsymbol{\theta(X \rightarrow c) \geq \theta min}$ .. (8) is generated.

The system also provide a feedback facility in case of those words that are not found in the database and are expected to be a spam, in such situations user can report such words to the administrator. The administrator is responsible for development and maintenance of database. The administrator would perform an update operation in the database if he finds the reported word or link to be a spam.

Conclusion:-

This paper reports a spam email detection method, where previous knowledge about spam emails is assumed in the form of set of links and words which are stored in the database. The motivation of this proposal of ours is due to the fact that spammers are persistently creating new strategies of spam emails, which may be completely different from the previous ones, to defeat the spam detection engines. In order for a spam detection engine to detection the new strains of spam emails, it has to operate without any knowledge about these new spam emails

Acknowledgement:-

Our sincere thanks to M. H. Saboo Siddik College of Engineering, Department of Computer Engineering, for giving us the initiative to do constructive work. We also thank anonymous reviewers for their constructive suggestions.

References:-

1. Adriano Veloso, Eager, Lazy and Hybrid Algorithms for Multi-Criteria Associative Classification, Universidade Federal de Minas Gerais(UFMG),Brazil,2005.
2. Pattern Recognition and Machine Learning, Christopher Bishop, Springer-Verlag, 2006.
3. Adriano Veloso, Lazy Associative Classification for Content-based Spam Detection,Fourth Latin American Web Congress-2006.
4. Konstantin Tretyakov, Machine Learning Techniques in Spam Filtering, Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004.
5. Masaru Takesue, Cascaded Simple Filters for Accurate and Lightweight Email-Spam Detection, Fourth International Conference on Emerging Security Information, Systems and Technologies,2010.
6. Adriano Veloso, Lazy Associative Classification, Sixth International Conference on Data Mining (ICDM'06),2006.
7. Geerthik. S, Survey on Internet Spam: Classification and Analysis, Int.J.Computer Technology & Applications,Vol 4 (3), IJCTA May-June 2013